Benchmarking Cognitive Biases in Large Language Models as Evaluators

Year: 2023 | *Citations: 123* | *Authors: Ryan Koo, Minhwa Lee, Vipul Raheja*

Abstract

Large Language Models are cognitively biased judges. Large Language Models (LLMs) have recently been shown to be effective as automatic evaluators with simple prompting and in-context learning. In this work, we assemble 15 LLMs of four different size ranges and evaluate their output responses by preference ranking from the other LLMs as evaluators, such as System Star is better than System Square. We then evaluate the quality of ranking outputs introducing the Cognitive Bias Benchmark for LLMs as Evaluators (CoBBLEr), a benchmark to measure six different cognitive biases in LLM evaluation outputs, such as the Egocentric bias where a model prefers to rank its own outputs highly in evaluation. We find that LLMs are biased text quality evaluators, exhibiting strong indications on our bias benchmark (average of 40% of comparisons across all models) within each of their evaluations that question their robustness as evaluators. Furthermore, we examine the correlation between human and machine preferences and calculate the average Rank-Biased Overlap (RBO) score to be 49.6%, indicating that machine preferences are misaligned with humans. According to our findings, LLMs may still be unable to be utilized for automatic annotation aligned with human preferences. Our project page is at: https://minnesotanlp.github.io/cobbler.